DOCUMENT RESUME

ED 408 325 TM 026 578

| | |
|---|---|
| AUTHOR | Wolfe, Edward W.; Chiu, Chris W. T. |
| TITLE | Measuring Change over Time with a Rasch Rating Scale Model. |
| SPONS AGENCY | American Coll. Testing Program, Iowa City, Iowa. |
| PUB DATE | Mar 97 |
| NOTE | 48p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997). |
| PUB TYPE | Reports - Evaluative (142) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Change; Item Response Theory; Measurement Techniques; *Portfolio Assessment; Portfolios (Background Materials); Probability; *Rating Scales; Teachers; *Time |
| IDENTIFIERS | Additive Models; Anchoring Devices; Calibration; FACETS Computer Program; Linear Models |

ABSTRACT

When measures are taken on the same individual over time, it is difficult to determine whether observed differences are the result of changes in the person or changes in other facets of the measurement situation (e.g. interpretation of items or use of rating scale). This paper describes a method for disentangling changes in persons from changes in Likert-type questionnaire items and rating scales. The procedure relies on anchoring strategies in Rasch measurement to create a common frame of reference for interpreting measures that are taken at different times. The Rasch Rating Scale Model is an additive linear model that describes the probability that a specific person will respond to a specific Likert-type item with a specific rating scale step. How to perform the procedure is illustrated using data from 168 teachers responding to a questionnaire about barriers to the implementation of a portfolio assessment. The five steps used to measure change over time are summarized as: (1) evaluate scale and item invariance; (2) correct the scale calibrations; (3) benchmark the time-1 estimates; (4) correct the time-2 person measures; and (5) correct the time-2 item calibrations. Applications of the procedure reveal two changes in teachers that were not apparent when the correction was not applied to the analyses. Five appendixes present the FACETS command files for each step. (Contains 2 figures, 5 tables, and 15 references.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

RUNNING HEAD: Measuring Change

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Edward Wolfe
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 408 325

# Measuring Change Over Time with a Rasch Rating Scale Model

Edward W. Wolfe

Educational Testing Service, Princeton, NJ

Chris W.T. Chiu

Michigan State University

## Author Note

Edward W. Wolfe, Center for Performance Assessment; Chris W.T. Chiu, Measurement and Quantitative.

This research was supported by ACT inc. and a post-doctoral fellowship at Educational Testing Service. The authors thank Carol Myford and Mark Reckase for their input on a previous version of this manuscript. This paper was presented at the *International Objective Measurement Workshop 9* in Chicago, IL (March, 1997).

Correspondence concerning this article should be addressed to Edward W. Wolfe, ETS Mail Stop 11-P, Princeton, New Jersey, 08541-0001. Electronic mail may be sent via Internet to ewolfe@ets.org.

BEST COPY AVAILABLE

TM026578

Abstract

When measures are taken on the same individual over time, it is difficult to determine whether observed differences are the result of changes in the person or changes in other facets of the measurement situation (e.g., interpretation of items or use of rating scale). This paper describes a method for disentangling changes in persons from changes in Likert-type questionnaire items and rating scales. The procedure relies on anchoring strategies in Rasch measurement to create a common frame of reference for interpreting measures that are taken at different times. An example of how to perform the procedure is given.

Measuring Change Over Time Using a Rasch Rating Scale Model

Measuring change over time presents particularly difficult problems for program evaluators, especially when the indicator of change is a psychological construct such as teacher attitudes or student achievement. A number of potential confounds may distort the measurement of change, making it unclear whether the observed changes in the outcome variable are due to the intervention or some other effect such as regression toward the mean (Lord, 1967), maturation of subjects, or idiosyncrasies of subjects who drop out of the program (Cook & Campbell, 1979). When rating scales or assessment instruments are used to measure changes in an outcome variable, additional potential confounds are introduced into the evaluation design. For example, subjects may improve their performance on an assessment instrument that is used as both a pre-test and post-test because of familiarity with the test items (Cook & Campbell, 1979). Alternatively, when changes are measured with Likert-type questionnaires, subjects may interpret the items or the rating scale options differently on the two occasions (Wright, 1996b).

This article describes an equating procedure that can be applied to rating scale data to compensate for the latter of these potential confounds to measuring change over time. That is, we describe a method for reducing the effect that changes in subjects' interpretations of questionnaire items and rating scale options may have on the measurement of change in the underlying construct. By making the proposed correction, evaluators can eliminate at least one threat to the valid interpretation of changes in attitudes or opinions as measured by Likert-type questionnaires. In this article, we outline the procedures for making this

correction, illustrate how these procedures are carried out, and demonstrate how the employment of these procedures can lead to the discovery of changes that would not be apparent otherwise.

## Theoretical Framework

In many program evaluation settings, evaluators are interested in measuring changes in the behaviors or attitudes of non-random samples of subjects who represent a population of interest. Changes in the measures of the outcome variable are typically inferred to be caused by participation in the program in question. Of course, numerous threats to the validity of this inference exist, and each of these threats highlights a potential confound that must be taken into account when designing an evaluation, collecting and analyzing data, and interpreting the results. These threats to the validity of the interpretations that are drawn from a program evaluation may relate to *statistical validity* (the accuracy of the statistical inferences drawn about the relationship between the program and the outcome variable), *construct validity* (the accuracy of the inferred relationship between the measurement procedures and the latent construct they are intended to represent), *external validity* (the accuracy of the inferred relationship between the subjects and the population that they are intended to represent), or *internal validity* (the accuracy of the theory-based inferences drawn about the relationship between the program and the outcome variable). Methods for avoiding or reducing each of these threats to drawing valid inferences are outlined by Cook and Campbell (1979).

The problem addressed by this article represents one of several potential threats to internal validity. That is, we are concerned with whether observed changes in the outcome variable are truly caused participation in the program or whether observed changes can be attributed to other variables that are byproducts of the evaluation setting. In a program evaluation, threats to internal validity may arise when changes in subjects can be attributed to developmental maturation, changes in subjects' familiarity with the measurement instrument, mortality of subjects, the procedures used to assign subjects to treatments, statistical regression toward the mean, or changes in the measurement instrument rather than the treatment itself. The threat to internal validity that we discuss arises when Likert-type questionnaire items are used to measure attitudinal changes. More specifically, we are concerned with the degree to which changes in the way people interpret questionnaire items and use rating scales confounds the measurement of changes in attitudes or opinions.

Prior measurement research in this area has shown that subjects' interpretations of items or uses of rating scales may change over time and that this is a common concern for those who use questionnaires to measure outcome variables. For example, Zhu (1996, April) investigated how children's psychomotoric self-efficacy changes over time. In this study, children completed a questionnaire designed to measure the strength of their confidence about their abilities to perform a variety of physical exercises. The results of this study indicate that some of the activities were perceived as being less difficult to perform, relative to the remaining activities, over repeated administrations of the questionnaire. Such differential

functioning of items over time threatens the validity of interpretations that might be drawn from the results of this study.

In order to evaluate changes in persons over time, the items and rating scales that are used to measure this change must be stable across multiple administrations of the questionnaire. Only if items and scales demonstrate such stability can differences between different measures of the persons be validly interpreted (Wilson, 1992; Wright, 1996b). To further exacerbate the problem, summated composite scores are not comparable across time when items are added, removed, or reworded; items are skipped by some subjects; or response options change from pre-test to post-test--all problems that are common when questionnaires are used (Roderick & Stone, 1996, April). Because of problems such as these, scaling methods are often used to place measures from different administrations of a questionnaire onto a common scale.

Rasch Measurement

Rasch measurement is a latent trait modeling technique that has proven useful for solving a variety of measurement problems. Applications of Rasch measurement to the analysis of questionnaire data result in several beneficial conditions. First, Rasch measurement places each facet of the measurement context (e.g., items and persons) on a common underlying linear scale. This results in linear measures that can be subjected to traditional statistical analyses while allowing for unambiguous substantive interpretations of the meaning of person performance as it relates to item functioning. Furthermore, Rasch measurement provides unambiguous statistics for evaluating changes in person measures, item

calibrations, and scale step calibrations that are obtained under different measurement conditions. These statistics provide a valuable method for examining changes in individuals that may be attributable to specific programs. These statistics are central to the method we describe here.

Second, Rasch measurement produces sample-free estimates of person ability and item difficulty. That is, procedures used to estimate ability and difficulty parameters remove the influence of sampling variability from scaled scores so that valid generalizations can be made beyond the current sample of persons or collection of items. This means that similar estimates of person ability will be realized regardless of which items are used to measure that ability and that similar estimates of item difficulty will be realized regardless of the persons relative to whom that difficulty is evaluated. In applied settings, this feature is useful because it allows a person's ability to be determined even if that person does not respond to all of the items on a test or rating scale. This feature also allows poorly-functioning items to be removed from the analyses or replaced with better ones during subsequent testing or for problematic rating scale levels (i.e., response options) to be combined with other scale levels. As a result, data that would otherwise be eliminated from the analyses can be recovered.

Third, Rasch measurement models provide a framework for predicting how persons will respond to different items that have known qualities. That is, Rasch procedures can be used to derive expected response patterns that can be used to evaluate the extent to which individual items, persons, or rating scale steps are behaving in ways that are inconsistent with the measurement model. As a result, the suitability of the model for the measurement context

as well as the validity of the measures of individuals can be evaluated by examining the fit between the observed data and the expected response patterns.

Rating Scale Model

The Rasch Rating Scale Model (Andrich, 1978) is an additive linear model that describes the probability that a specific person (*n*) will respond to a specific Likert-type item (*i*) with a specific rating scale step (*x*). The mathematical model for this probability (Equation 1) describes this relationship in terms of a logistic odds ratio that contains three parameters: the person's *ability* ($\beta_n$), the item's *difficulty* ($\delta_i$), and the difficulty of each scale step (i.e., the threshold between two adjacent scale levels) ($\tau_x$). Calibration of questionnaire data results in a separate parameter estimate and a standard error for that estimate for each person, item, and scale step in the measurement context.

$$P(X_{ni}=x)=\frac{exp\sum_{j=0}^{x}\left[\beta_n-\left(\delta_i+\tau_j\right)\right]}{\sum_{x=0}^{m}exp\sum_{j=0}^{x}\left[\beta_n-\left(\delta_i+\tau_j\right)\right]}, \qquad x=0,1,...,m \qquad (1)$$

where, $P(X_{ni}=x)$ is the probability that a person *n* responds with rating scale category *x* to item *i*, which has *m+1* response options.

This model assumes that a common scale structure applies to each item (i.e., that $\tau_j$ is constant across items). The model also assumes that the data conform to the predictions of the Rating Scale Model. Equation 2 can be used to generate expected values for each combination of items and persons, and departures in the data from these expected values

indicate potentially misfitting items and persons. To evaluate the degree to which items and persons do not fit the Rating Scale Model, two statistics are generated for each parameter estimate (Wright & Masters, 1982). These fit statistics indicate the degree to which individual persons and items are behaving in ways that are inconsistent with the Rating Scale Model. Both fit statistics are based on the mean of the squared standardized residuals of the observed scores from their expected scores. The OUTFIT statistic is simply the mean of these standardized residuals. The INFIT statistic, on the other hand, weights each standardized residual so that item-person combinations that are well-matched (i.e., $\beta_n$ and $\delta_i$ are of similar magnitude) make a greater contribution to the magnitude of the fit statistic. As a result, OUTFIT statistics are more sensitive to unexpected responses to items that are not well-matched to the person's ability (i.e., are too difficult or too easy), while INFIT statistics are more sensitive to unexpected responses to items that are well-matched to the person's ability. The INFIT and OUTFIT statistics have an expected value of 1.00 and can range from 0.00 to ∞.

$$E_{ni} = \sum_{x=0}^{p} xP(X_{ni} = x), \qquad x=0,1,...p \qquad (2)$$

where, $E_{ni}$ is the expected response of examinee $n$ to item $I$, which has $p$ response options.

An important feature of the Rating Scale Model is that it allows one to evaluate the extent to which item calibrations are stable across samples of persons or the extent to which person measures are stable across samples of items (i.e., to determine the *invariance* of

parameter estimates). This feature is useful when comparing two groups of persons who respond to the same set of items or equating two tests (each composed of different items) that are taken separately by one group of examinees. In the present context, invariance evaluation is useful because it allows one to determine the extent to which item calibrations and person measures are stable across two measurement occasions. The stability of two parameter estimates ($\theta_1$ and $\theta_2$) that are obtained on different occasions is evaluated by examining the standardized difference (Equation 3) between the two estimates. The standardized differences for a population or item pool that conform to the Rating Scale Model have an expected value of 0.00 and an expected standard deviation of 1.00. Large departures in observed data from these expected values indicate parameters that are more or less stable over time than would be expected.

$$z = \frac{\theta_1 - \theta_2}{\sqrt{[SE(\theta_1)]^2 * [SE(\theta_2)]^2}} \qquad (3)$$

Examining Change Over Time with the Rating Scale Model

Measuring change over time requires a stable frame of reference, and differential functioning of items and rating scales disrupts the establishment of such a frame of reference. In order to measure changes in the performance of persons across time, other changes in the measurement framework must be eliminated or controlled. There are several methods for accomplishing this (Wright, 1996a). For example, facets other than the persons may be assumed to be constant by forcing the elements of each facet to remain fixed (e.g., by anchoring them to a common value). Alternatively, facets that exhibit noticeable change from

one occasion to another may be assumed to be truly different indicators of the construct in question and may, therefore, be treated as being completely different elements at each time. Finally, a compromise can be achieved between different administrations of an instrument by creating an "average" frame of reference and allowing facets to vary about that average.

The method we describe was originally proposed by Wright (1996b), and this method creates a common frame of reference by assuming that some elements of the measurement situation remain constant and by allowing others to vary over time. Many researchers desire to identify whether differences demonstrated by specific items or persons are large enough to be of importance, and the method presented in this article allows for such a distinction. Once a common frame of reference has been created, differences between person measures or between item calibrations at each measurement occasion can be evaluated by examining the standardized differences of the parameter estimates produced for each occasion. The method is described here as a five step procedure as portrayed in Figure 1.

Insert Figure 1 about here.

Step 1: Evaluate Scale and Item Invariance

The first step in using the Rating Scale Model to measure change over time is to determine whether interpretations of the scale steps and the items are stable across the two measurement occasions. If the item and step calibrations do demonstrate stability over time (i.e., they are invariant), then differences between person measures at the two occasions are valid indicators of changes in persons over time (i.e., they are free from potential confounding due to changes in interpretations of items or uses of rating scales). If the scale step and item

calibrations are not invariant over time, then the researcher must disentangle the changes in the scale steps, items, and persons to determine which elements of the measurement context are indeed changing (*Steps 2* through *5*).

To determine whether the scale step or item calibrations are invariant over time, one must generate two data sets--one containing the responses of each person ($n$) to each item ($i$) at Time 1 and the other containing the responses of each person to each item at Time 2. The layout of these data sets is shown in Figure 2. Item and step calibrations, as well as person measures, are obtained for each data set separately so that there is a pair of estimates, one for Time 1 and one for Time 2, for each scale step (F1.1 & F1.2), each item (D1.1 & D1.2), and each person (B1.1 & B1.2) in the study (where B1.1 refers to the person measures from *Step 1* for Time 1 and B1.2 refers to the person measures from *Step 1* for Time 2).

Insert Figure 2 about here

To evaluate the invariance of item and step calibrations over time, one compares the pair of calibrations for each element of these two facets. That is, one compares F1.1 and F1.2 for each step level and compares D1.1 to D1.2 for each item. This comparison can be made using the standardized difference of the two estimates (Equation 3). Items or scale steps that exhibit large differences between their Time 1 and Time 2 calibrations (e.g., $|z| > 2.00$) are not invariant over time (i.e., they are unstable). Such differences between the way that the scale steps were used or the items were interpreted at each occasion may confound any inferences that are drawn based on observed differences in the person measures for Time 1 and Time 2, and the researcher must make corrections (*Steps 2* through *5*). If there are no large differences

between step and item calibrations from the two occasions, then it is safe to interpret the differences between the person measures from the two occasions as indicators of change in persons over time. Again, this can be done by examining the standardized differences (Equation 3) between the two measures for each person. (B1.1 & B1.2).

Step 2: Correct the Scale Calibrations

If the analyses in *Step 1* reveal that the step or item calibrations are not stable across time, then there is a need to constrain this variability before interpreting observed changes in person measures because this interpretation cannot be valid unless a frame of reference is created that links Time 1 and Time 2. As suggested previously, there are two anchoring methods that can be used to create a common frame of reference (e.g., assuming a constant calibration value or adopting an average calibration value). There are also multiple components of the measurement context that can be anchored to create the frame of reference (e.g., scale steps or item calibrations). The current method adopts an average value for the scale steps to create the measurement framework. Thus, this method assumes that a common underlying, equal-interval scale adequately portrays the data and that departures from that underlying scale are due only to random fluctuations.

Therefore, the second step in measuring change over time is to correct the step calibrations so that person measures and item calibrations from Time 1 and Time 2 can be compared on a common underlying rating scale. To accomplish this, persons are assumed to vary from Time 1 to Time 2, and items are assumed to be invariant from Time 1 to Time 2. That is, persons are treated as being different objects of measurement on each of the two

occasions, and items are treated as being the same objects at each of the two occasions. This means that the two data sets from *Step 1* must be reconfigured by assigning two unique identifiers to each person--one for Time 1 responses (*n*.1) and one for Time 2 responses (*n*.2)--and appending them (i.e., stacking them to create a single data set). The format of the reconfigured data is shown in Figure 2.

Insert Figure 3 about here

This stacked data set is analyzed to obtain step calibrations that are consistent with person performance and item functioning across both occasions. The values of these common-scale estimates (F2.1&2) are used in *Steps* 3 through *5* as anchors for the scale steps. Analysis of the stacked data set also produces a single set of item calibrations (D2.1&2) and two separate measures for each person--one portraying the person at Time 1 (B2.1) and another portraying the person (as a different person) at Time 2 (B2.2). These item calibrations and person measures are ignored.

Step 3: Benchmark the Time 1 Estimates

Once a common rating scale has been created for the two occasions, that scale is used as a frame of reference for the Time 1 and Time 2 data sets. In *Step* 3 of the procedure, the Time 1 data are re-analyzed using the step calibrations from *Step 2* (i.e., F2.1&2) as anchors for that facet. This results in three sets of estimates: 1) benchmark item calibrations for items that were found to be invariant across time in *Step 1* (D3.1*), 2) item calibrations for items that were found to be unstable over time in *Step 1* (D3.1'), and 3) person measures for all persons (B3.1). Each of these sets of estimates is referenced to the common scale that was

created in the *Step 2* analyses. Each set of estimates is also used as the basis for measuring change in *Steps 4* and *5*.

Step 4: Correct the Time 2 Person Measures

In *Steps 2* and *3*, a frame of reference was created for interpreting changes in person measures at Time 2 by creating a rating scale that is common to both occasions and determining the calibrations for items that are invariant across time. In *Step 4*, the Time 2 data are re-analyzed by anchoring the steps on the common-scale values obtained in *Step 2* (i.e., F2.1&2) and anchoring the invariant items from *Step 1* on the benchmark item calibrations from *Step 3* (D3.1*). The items that were found to be unstable from one occasion to the next, however, are not anchored (i.e., they are allowed to float). This means that these items are treated as if they are different items on each occasion, and interpretations of change will have to be made in light of this fact.

The *Step 4* analyses produce person measures (B4.2) that are referenced to a rating scale that is valid for both Time 1 and Time 2 and a set of items that are invariant across time. Any differences between these corrected person measures (B4.2) and the benchmark measures obtained in *Step 3* (B3.1) indicate changes in persons, rather than interpretations of items or uses of the rating scale, over time. For each person, the benchmark Time 1 measure (B3.1) and the corrected Time 2 measure (B4.2) can be compared using the standardized difference as shown in Equation 3. Persons that exhibit large variability (e.g., $|z| > 2.00$) have changed over time. The analysis also produces calibrations for the unstable items (i.e., the items that were allowed to float--D4.2'). These calibrations may be ignored.

Step 5: Correct the Time 2 Item Calibrations

The final step in the procedure is to determine the extent to which item functioning changed over time while controlling for changes in person measures. In *Step 5*, the Time 2 data are re-calibrated by anchoring the scale steps on the joint calibrations obtained in *Step 2* (F2.1&2) and anchoring the person measures on the corrected Time 2 estimates from *Step 4* (B4.2). All items are allowed to float. This analysis results in item calibrations (for all items) at Time 2 (D5.2) that are corrected for changes in both the interpretation of the rating scale and the performance of people. To determine how much item functioning changed across occasions, the benchmark Time 1 item calibrations (D3.1* and D3.1') are compared to the corrected Time 2 item calibrations (D5.2). The comparison can be made by computing the standardized differences between these two estimates (Equation 3). This comparison is free from potential confounds due to changes in the use of the rating scale or the performance of persons across time. It is important to note that calibrations for items that were found to be unstable over time in the *Step 1* analyses have been treated as different items in the estimation of person measures regardless of how much their corrected calibrations differ.

Example

The remainder of this article illustrates how this procedure can be applied to questionnaire data. We demonstrate this technique on data that are typical of many program evaluations (i.e., *pre-test, intervention, post-test*). Our analyses emphasize how using the procedure results in different interpretations of how persons and items change over time.

Subjects

The data for our demonstration come from mathematics, science, and language arts teachers from 14 public and private secondary schools in different regions of the United States. These teachers participated in a nine-month program designed to help them develop portfolio assessments. Approximately 12 teachers from each school participated in the program ($n$=168). At the beginning of the school year (in September), teachers responded to a questionnaire designed to assess the strength with which teachers perceive potential barriers to the implementation of a portfolio assessment program to be problematic (Wolfe & Miller, in press). After participating in the program for an academic year (in June), teachers completed the questionnaire a second time. A comparison of a teacher's responses from September (Time 1) with the responses provided in June (Time 2) was interpreted as a measure of change in the teacher's perception of barriers to the implementation of portfolio assessments. Fairly complete data for Time 1 and Time 2 were available for 117 of the 168 teachers who participated in the program (a 30% attrition rate).

Instrument

The questionnaire asked teachers how problematic they perceived 30 potential barriers to the implementation of a portfolio assessment system to be. The barriers referenced issues such as the amount of *time* required to use portfolios, resistance from *people* to the idea of using portfolios, the difficulty of assigning *scores* to portfolio entries, changes in *instruction* that are required when portfolios are used, and the availability of *resources* for using portfolio assessment. Each barrier was formatted as the stem for a four-point Likert-

type item. Teachers responded to each barrier by indicating whether the barrier is a(n) *unlikely*, *minor*, *difficult*, or *serious* problem. For each of the 30 barriers, teachers indicated the option that best describes the difficulty of that specific barrier. *Unlikely* problems were defined as those that would likely have no impact on the teacher's use of portfolios. *Minor* problems were those that may cause the teacher to use portfolios differently than they would be used in an ideal situation. *Difficult* problems were defined as problems that may cause the teacher to reconsider using portfolios in his or her classroom. *Serious* problems were those that would cause the teacher not to use portfolios at all.

## Analyses and Results

These data were analyzed with a Rasch Rating Scale Model. For substantive meaning, all facets were scaled so that higher logit values are associated with more difficult portfolio implementation. That is, higher values of teacher measures are associated with the perception of portfolio implementation as being more difficult, and higher values of barrier and rating scale step calibrations are associated with barriers that are more difficult to overcome. In each of the following sections, we detail the steps of the anchoring method described by Wright (1996b). Prior to illustrating the five steps, however, the fit of the data to the model is evaluated because a necessary prerequisite for interpreting the results of Rating Scale Model analyses is to verify that the data can be adequately described by the model.

### Evaluating Fit

For *Step 1*, the data are placed in two data sets--one data set containing the teachers' responses from September (Time 1) and the other containing teachers' responses from June

(Time 2) (see Figure 1). Each data set contains three variables: 1) teacher (person) identifier, 2) barrier (item) identifier, and 3) the teacher's response (rating) to that barrier. In our description, we use FACETS (Linacre, 1989), a computer program designed to carry out multi-faceted Rasch analyses, to obtain parameter estimates for these data sets. It should be noted, however, that these analyses can be performed using any item response software that allows for the analysis of rating scale data and the anchoring of measurement facets. The two data sets, one from each of the two occasions, are calibrated on separate FACETS analyses. An example command file for performing *Step 1* with FACETS is shown in Appendix A. A similar command file is written for the June data. These analyses result in one set of calibrations for the September data and a different set of calibrations for the June data.

To evaluate the fit of the data to the model, and hence the appropriateness of using the Rating Scale Model, the fit statistics for the parameter estimates of each teacher, barrier, and rating scale step must be examined at each occasion to determine whether the data conform to the requirements of the Rasch Rating Scale Model. The INFIT and OUTFIT mean square residuals have a mean of 1.00 and a range from 0.00 to $\infty$, and values between 0.6 and 1.5 are considered to be within an acceptable range (Engelhard, 1994; Lunz, Wright, & Linacre, 1990; Wright & Linacre, 1994). One of the barrier in our questionnaire had large fit statistics for both the September and June data, so that barrier was discarded from both data sets leaving a total of 29 barriers. Six of the teachers showed poor fit to the model in both September and June, and another five responded with the lowest level of the score scale to all

barriers on both occasions (and hence could not adequately be measured with the barriers in the questionnaire). After removing these 11 teachers from both data sets, the number of teachers was 106. None of the step calibrations showed poor fit to the Rating Scale Model, so the rating scale levels were not altered.

Step 1: Evaluate Rating Scale and Barrier Invariance

As described in the previous section, the September and June responses were analyzed separately so that each teacher, barrier, and rating scale step received a pair of parameter estimates--one for September and one for June. The pair of estimates for each teacher, barrier, and rating scale step are referred to here as B1.1 and B1.2, D1.1 and D1.2, and F1.1 and F1.2, respectively. To determine whether differences between B1.1 and B1.2 are valid indicators of change in teacher measures over time, we computed the standardized differences (Equation 3) between each pair of step calibrations (F1.1 and F1.2) and each pair of barrier calibrations (D1.1 and D1.2). The parameter estimates for September and June, their standard errors, and the associated standardized differences are shown in Tables 1 and 2.

Insert Table 1 about here

Insert Table 2 about here

The analyses from *Step 1* reveal that there are large differences in the way that the rating scale steps were used at September and June as indicated by the large standardized difference for two of the three scale step calibrations (Table 1). Furthermore, several of the barriers showed unexpectedly large changes in their calibrations over the two administrations of the questionnaire. In fact, 7 of the 29 barriers (24%) have absolute standardized differences

greater than 2.00. This is a large percentage of barriers when compared to the expectation derived from the standard normal distribution (about five percent). The broad distribution of values results in a standard deviation of *z* scores (1.82) that is considerably larger than the expected value of 1.00. These statistics suggest that differences in the functioning of barriers and rating scale steps over time may cloud any interpretations that we make of differences in teacher measures, so our example proceeds with *Steps 2* through *5* of Wright's (1996b) procedure.

Step 2: Correct the Scale Calibrations

In *Step 2* of the procedure, a common rating scale is created so that teacher attitudes and barrier severity estimates from September and June can be estimated in a common frame of reference. To this end, we stack the two data sets from *Step 1*, reassigning teacher identifiers to each teacher for the June responses. In our example, we simply added 1000 to the original identifier (as shown in the example FACETS command file presented in Appendix B). Because this step of the analysis portrays each teacher as being a different person in June than in September and allows barriers to remain stable across administration of the questionnaire, the output of this command file results in a pair of measures for each teacher and a single calibration for each barrier. All of these values are ignored. The rating scale step calibrations (F2.1&2) from this analysis, however, are of interest and will be utilized as anchor values for the remaining steps of the procedure. Table 3 compares the scale step calibrations from *Step 1* of the procedure to those obtained from *Step 2*. As one would expect, the values from *Step 2* (i.e., the step calibrations for the scale that is common to

September and June) are between the two values obtained in *Step 1* (i.e., the step calibrations for the separate September and June scales).

Insert Table 3 about here

Step 3: Benchmark September Estimates

In *Step 3*, benchmark estimates are obtained for teachers (B3.1) and barriers (D3.1) by anchoring rating scale steps on the values obtained in *Step 2* (i.e., F2.1&2). Appendix C shows an example FACETS command file for this analysis. Note that the command file is the same as the command file used in *Step 1* with the exception that rating scale steps are now anchored on their F2.1&2 values. The data file is the same one used for the September analysis in *Step 1*. The *Step 3* analyses result in three sets of values for the September data. The benchmark teacher measures (B3.1), the benchmark calibrations for the barriers that were found to be invariant in *Step 1* (D3.1*), and the calibrations for the barriers that were found to be unstable in *Step 1* (D3.1') are all used as the basis for measuring changes in teachers and barriers in *Steps 4* and *5*.

Step 4: Correct the June Teacher Measures

In *Step 4*, the common rating scale step calibrations from *Step 2* (F2.1&2) and the benchmark barrier calibrations obtained in *Step* 3 for the 22 items that were found to be invariant across time (D3.1*) are used as anchors so that corrected teacher measures can be estimated for the June data. As shown in Appendix D, the seven barriers that were found to be unstable across time in *Step 1* are not anchored (i.e., they are allowed to float). Note that new calibrations (D4.2') are obtained for these barriers in *Step 4*, but these values are ignored.

Otherwise, the procedures for analyzing the June data are the same as they were in *Step 1*. The resulting teacher measures (B4.2) have been corrected for changes in perceptions of barriers and uses of the rating scale over time through this anchoring process. As a result, a comparison of the corrected June teacher measures (B4.2) with the benchmark September teacher measures (B3.1) reveals how people have changed over time without any confounding from changes in barrier or rating scale functioning. This comparison can be made by examining the standardized difference (Equation 3) for each teacher's pair of corrected measures.

Table 4 shows the standardized difference of measures for the 20 teachers with the largest absolute change in $z$ values between their uncorrected (i.e., *Step 1*) and corrected (i.e., *Steps 3* and *4*) estimates. That is, these teachers showed the most dramatic change in their positions on the underlying scale when their September and June measures were corrected for changes in barrier perception and rating scale functioning over time. Note that if changes in individual teachers are declared to be large when $|z|>2.00$, then decisions concerning the significance of observed change would be different for 3 of the 20 teachers shown in Table 4, depending on whether corrected or uncorrected $z$ values are used. That is, *Teachers 4* ($z_{uncorrected}$=-2.78 and $z_{corrected}$=-1.98) and *9* ($z_{uncorrected}$=-2.03 and $z_{corrected}$=-1.29) would seem to have changed only if corrected $z$ values were examined, while *Teacher 20* ($z_{uncorrected}$=-1.74 and $z_{corrected}$=-2.42) would seem to have changed only if uncorrected $z$ values were examined. In the total sample ($N$=106), 11% of the teachers' standardized differences showed such a discrepancy between their corrected and uncorrected $z$ values. These differences at the

individual teacher level suggest that our procedures have removed error that confounds the internal validity of our study.

Examination of the descriptive statistics of the corrected and uncorrected standardized differences for the entire sample of teachers illuminates what this means for the overall measurement of change. In the last row of Table 4, note that the mean of the corrected standardized differences is smaller than the mean of the uncorrected values. This means that, by removing the influence of changes in the perception of barriers and changes in the use of the rating scale, we have reduced the apparent change in teachers over time. If we use a *t* test to examine the size of the change over time, we notice that the corrected teacher measures are somewhat more similar than are the uncorrected teacher measures [$t(105)=-2.35$, $p=.0105$, $\eta^2=.01$ and $t(105)=-3.93$, $p=.0001$, $\eta^2=.04$, respectively].[1] Although the overall strength of this relationship is not great in either case (i.e., $h^2$ is small), the difference between the two statistics is quite large. As a result, it seems that there is less variability within teachers over time when measures are corrected. This fact emphasizes that the removal of the confounding influence of changes in the perception of barriers and uses of the rating scale results in different interpretations of changes in teachers.

Insert Table 4 about here

Step 5: Correct the June Barrier Calibrations

In *Step 5*, the common rating scale step calibrations from *Step 2* (F2.1&2) and the corrected person measures for June obtained in *Step 4* (B4.2) are used as anchors so that

corrected barrier calibrations can be estimated for the June data. As shown in the example command file in Appendix E, this anchoring is the only difference between the analyses of the June data for *Steps 1* and *5*. The resulting barrier calibrations (D5.2) for all items have been corrected for changes in teachers and uses of the rating scale over time. The corrected June barrier calibrations can be compared to the benchmark calibrations for September (D3.1) to identify how the perception of barriers changed over time. As in the previous analyses, this comparison is made by examining the standardized difference (Equation 3) for each barrier's pair of corrected calibrations.

Table 5 shows the corrected and uncorrected standardized difference of the calibrations for each barrier. The most interesting feature of these statistics concerns the conclusions that would be drawn about how the perception of barriers changes over time depending on whether one examines the corrected or the uncorrected standardized differences. A few of the barriers exhibit the same tendencies regardless of which set of values is examined. For example, the barriers associated with *Resources 2* and *6*, *People 5*, and *Scores 2* all seem to be less problematic at the beginning of the year (i.e., they have positive standardized differences) while the barriers associated with *People 1* and *6* seem to be more problematic at the end of the year (i.e., they have negative standardized differences) regardless of whether corrected or uncorrected barrier calibrations are examined. On the other hand, our conclusions about other barriers would change depending on whether we examine the corrected or the uncorrected calibrations. If we consult the uncorrected standardized differences, we would conclude that the barrier associated with *People 2* became more

problematic and that the barrier associated with *Instruction 4* became less problematic with time. On the other hand, if we examine the corrected standardized differences, we would conclude that the perception of neither of these barriers changed a large amount. Instead, we would conclude that the barriers associated with *Time 2* and *3* and *Instruction 2* each became more problematic with time and that the barrier associated with *Resources 4* was perceived as being less problematic over time. Depending on whether use the corrected or uncorrected barrier calibrations, we may draw very different conclusions about ways that participation in this program influences which of these barriers to portfolio implementation are problematic for teachers.

Insert Table 5 about here

## Conclusions

We have demonstrated a procedure for removing potentially-confounding sources of variability from the measures of changes in persons over time. Application of this procedure to the data in our example revealed two changes in the teachers that were not apparent when this correction was not applied to the analyses. First, use of this procedure revealed differences in the degree to which measured changes in individual teachers were large enough to be considered meaningful. That is, depending on whether teacher measures were corrected or not, different conclusions would have been drawn concerning which teachers' perceptions of barriers indeed changed as they participated in the program. Thus, when individual teachers who have special characteristics are of interest, the procedure we illustrated may lead to different conclusions concerning the influence of a program on these participants. Second,

our procedure revealed differences in the degree to which teachers as a group changed over time. By removing the confounding influence of changes in the perception of barriers and the way that the rating scale was used, we were able to show that teachers did not change as much as was implied by the uncorrected teacher measures. This procedure also revealed changes in the structure of the underlying variable that resulted from disentangling the various changes that took place in the measurement context. By removing changes in rating scale use and changes in teacher beliefs from the barrier calibrations, we were able to detect several changes in barrier functioning that were not apparent prior to making this correction.

Overall, this procedure seems useful for disentangling changes in the item functioning and rating scale use from changes in person performance when likert-type questionnaires are used to measure the impact that a program has in participants. As a result, the procedure could prove useful for program evaluators who are interested in measuring changes in attitudes and opinions. Further exploration of this procedure might focus on whether the method can be adapted to multi-faceted measurement contexts or to measurement models based on different response structures (e.g., partial credit models). Extending this correction procedure to settings with more than two measurement occasions would also be a useful application.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Cook, T.C., & Campbell, D.T. (1979). *Quasi-Experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Engelhard, G.J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Lunz, M.E., Wright, B.D., & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Lord, F.M. (1967). Elementary models for measuring change. In C.W. Harris, *Problems in Measuring Change* (pp. 21-38). Madison, WI: University of Wisconsin Press.

Roderick, M., & Stone, S. (1996, April). Is it changing opinions or changing kids? Manuscript presented at the 1996 Annual Meeting of the American Educational Research Association, New York, NY.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 6*, 334-349.

Wilson, M. (1992). Measuring changes in the quality of school life. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 77-96). Norwood, NJ: Ablex Publishing.

Wright, B.D. (1996a). Comparisons require stability. *Rasch Measurement Transactions , 10,* 506.

Wright, B.D. (1996b). Time 1 to Time 2 Comparison. *Rasch Measurement Transactions , 10,* 478-479.

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 170.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wolfe, E.W., & Miller, T.R. (in press). Barriers to the implementation of portfolio assessment in secondary education. *Applied Measurement in Education.*

Zhu, W. (1996, April). Many-Faceted Rasch Analysis of Children's Change in Self-Efficacy. Manuscript presented at the 1996 Annual Meeting of the American Educational Research Association, New York, NY.

Appendix A. FACETS Command File for *Step 1*

```
title = STEP 1, TIME 1: EVALUATE SCALE AND BARRIER INVARIANCE
output = STEP1_T1.OUT,STEP1_T1.ANC
facets = 2
models =
?,?,likert,1
*
Rating Scale=LIKERT,R3,general
0=UNLIKELY
1=MINOR
2=DIFFICULT
3=SERIOUS
*
labels =
   1,TEACHER
      1 = 1; teacher 1 time 1
      2 = 2; teacher 2 time 1
      .
      .
      .
      106 = 106; person 106 at time 1
*
   2,BARRIER
      1 = A; barrier A
      2 = B; barrier B
      .
      .
      .
      29 = CC; barrier CC
data = STEP1_T1.DAT; Time 1 data configured as shown in Figure 1
```

## Appendix B. FACETS Command File for *Step 2*

```
title = STEP 2, TIME 1&2: CORRECTED SCALE CALIBRATIONS
output = STEP2.OUT,STEP2.ANC
facets = 2
models =
?,?,likert,1
*
Rating Scale=LIKERT,R3,general
0=UNLIKELY
1=MINOR
2=DIFFICULT
3=SERIOUS
*
labels =
   1,TEACHER
      1 = 1; teacher 1 time 1
      2 = 2; teacher 2 time 1
      .
      .
      .
      106 = 106; teacher 106 time 1
      1001 = 1001; teacher 1 time 1
      1002 = 1002; teacher 2 time 1
      .
      .
      .
      1106 = 1106; teacher 106 time 1
*
   2,BARRIER
      1 = A; barrier A
      2 = B; barrier B
      .
      .
      .
      29 = CC; barrier CC
data = STEP2.DAT; Time 1 & Time 2 data configured as shown in Figure 2
```

## Appendix C. FACETS Command File for *Step 3*

```
title = STEP 3, TIME 1: BENCHMARK TIME 1 ESTIMATES
output = STEP3.OUT,STEP3.ANC
facets = 2
models =
?,?,likert,1
*
Rating Scale=LIKERT,R3,general
0=UNLIKELY,0,A; always anchor on 0
1=MINOR,-1.685754,A; anchor on value from Step 2
2=DIFFICULT,.088102,A; anchor on value from Step 2
3=SERIOUS,1.597652,A; anchor on value from Step 2
*
labels =
   1,TEACHER
      1 = 1; teacher 1 time 1
      2 = 2; teacher 2 time 1
      .
      .
      .
      106 = 106; teacher 106 at time 1
*
   2,BARRIER
      1 = A; barrier A
      2 = B; barrier B
      .
      .
      .
      29 = CC; barrier CC
data = STEP1_T1.DAT; Time 1 data configured as shown in Figure 1
```

Appendix D. FACETS Command File for *Step 4*

```
title = STEP 4, TIME 2: CORRECTED TIME 2 PERSON MEASURES
output = STEP4.OUT,STEP4.ANC
facets = 2
models =
?,?,likert,1
*
Rating Scale=LIKERT,R3,general
0=UNLIKELY,0,A; always anchor on 0
1=MINOR,-1.685754,A; anchor on value from Step 2
2=DIFFICULT,.088102,A; anchor on value from Step 2
3=SERIOUS,1.597652,A; anchor on value from Step 2
*
labels =
   1,TEACHER
      1001 = 1001; teacher 1 time 2
      1002 = 1002; teacher 2 time 2
      .
      .
      .
      1106 = 1106; teacher 106 time 2
*
   2,BARRIER,A
      1 = A,-.4824974; barrier A (invariant--anchor on Step 3 value)
      2 = B; barrier B (unstable--allow to float)
      .
      .
      .
      29 = CC,-.600314; barrier CC (invariant--anchor on Step 3 value)
data = STEP1_T2.DAT; Time 2 data configured as shown in Figure 1
```

Appendix E. FACETS Command File for *Step 5*

```
title = STEP 5, TIME 2: CORRECTED TIME 2 ITEM CALIBRATIONS
output = STEP5.OUT,STEP5.ANC
facets = 2
models =
?,?,likert,1
*
Rating Scale=LIKERT,R3,general
0=UNLIKELY,0,A; always anchor on 0
1=MINOR,-1.685754,A; anchor on value from Step 2
2=DIFFICULT,.088102,A; anchor on value from Step 2
3=SERIOUS,1.597652,A; anchor on value from Step 2
*
labels =
   1,TEACHER,A
      1001 = 1001,-2.20642; teacher 1 time 2 (anchor on Step 3 value)
      1002 = 1002,-1.066148; teacher 2 time 2 (anchor on Step 3 value)
      .
      .
      .
      1106 = 1106,-0.054304; teacher 106 time 2 (anchor on Step 3 value)
*
   2,BARRIER
      1 = A; barrier A
      2 = B; barrier B
      .
      .
      .
      29 = CC; barrier CC
data = STEP1_T2.DAT; Time 2 data configured as shown in Figure 1
```

## Footnotes

[1]$\eta^2$ (eta-squared) is a measure of association analogous to $R^2$ in regression analysis. $\eta^2$ indicates the proportion of total variance that is accounted for by the variance associated with the treatment [i.e., $SS_{explained}/SS_{total}$ (Snyder & Lawson, 1993)]. In the text, only 1% of the total variance is accounted for by variability over time when corrected measures are compared. However, 4% of the variance is accounted for by variability over time when the uncorrected measures are compared.

Table 1

Rating Scale Step Calibrations from *Step 1* for September and June

| Scale Step | F1.1 Logit | F1.1 Error | F1.2 Logit | F1.2 Error | z |
|---|---|---|---|---|---|
| Unlikely to Minor | -2.05 | 0.05 | -1.45 | 0.05 | -8.45 |
| Minor to Difficult | 0.04 | 0.05 | 0.10 | 0.05 | -0.77 |
| Difficult to Serious | 2.01 | 0.11 | 1.36 | 0.09 | 4.59 |
| **Mean** | **0.00** | **0.07** | **0.00** | **0.06** | **-1.55** |
| **(SD)** | **(2.03)** | **(0.03)** | **(1.41)** | **(0.02)** | **(6.56)** |

Note: F1.1 represents the rating scale step calibrations obtained in *Step 1* for September, and F1.2 represents the scale step calibrations obtained in *Step 1* for June. $|z|>2.00$ is considered large enough to indicate unstable uses of rating scale steps across occasions.

Table 2

Barrier Calibrations from *Step 1* for September and June

| Barrier | D1.1 Logit | D1.1 Error | D1.2 Logit | D1.2 Error | z |
|---|---|---|---|---|---|
| Instruction 1 | 0.49 | 0.16 | 0.47 | 0.14 | 0.08 |
| Instruction 2 | 0.16 | 0.16 | 0.51 | 0.13 | -1.69 |
| Instruction 3 | -0.01 | 0.16 | 0.18 | 0.14 | -0.94 |
| Instruction 4 | -0.46 | 0.17 | -0.01 | 0.14 | -2.04 |
| Instruction 5 | 0.25 | 0.16 | 0.50 | 0.13 | -1.20 |
| Instruction 6 | -0.68 | 0.18 | -0.38 | 0.15 | -1.27 |
| People 1 | -1.51 | 0.20 | -0.75 | 0.15 | -3.05 |
| People 2 | -1.07 | 0.18 | -0.62 | 0.15 | -1.94 |
| People 3 | 0.00 | 0.16 | 0.20 | 0.13 | -0.97 |
| People 4 | 0.05 | 0.16 | 0.27 | 0.14 | -1.03 |
| People 5 | -0.49 | 0.17 | -1.70 | 0.20 | 4.62 |
| People 6 | -0.41 | 0.17 | 0.06 | 0.14 | -2.11 |
| People 7 | -1.84 | 0.22 | -1.83 | 0.21 | -0.06 |
| People 8 | -1.67 | 0.21 | -1.93 | 0.22 | 0.84 |
| Resource 6 | 0.72 | 0.15 | 0.24 | 0.14 | 2.38 |
| Resources 1 | -0.55 | 0.17 | -0.41 | 0.14 | -0.62 |
| Resources 2 | 0.56 | 0.15 | 0.05 | 0.13 | 2.57 |
| Resources 3 | 0.40 | 0.15 | 0.22 | 0.13 | 0.93 |

Table continued.

Table 2 continued.

| | | | | | |
|---|---|---|---|---|---|
| Resources 4 | -0.16 | 0.16 | -0.54 | 0.15 | 1.73 |
| Resources 5 | -0.06 | 0.16 | -0.33 | 0.14 | 1.27 |
| Scores 1 | 0.66 | 0.15 | 0.46 | 0.13 | 0.98 |
| Scores 2 | 0.51 | 0.16 | -0.24 | 0.15 | 3.40 |
| Scores 3 | -0.05 | 0.16 | 0.01 | 0.14 | -0.28 |
| Scores 4 | 0.50 | 0.16 | 0.67 | 0.13 | -0.83 |
| Scores 5 | 0.49 | 0.16 | 0.59 | 0.13 | -0.46 |
| Scores 6 | 0.74 | 0.16 | 0.59 | 0.14 | 0.66 |
| Time 1 | 1.46 | 0.15 | 1.07 | 0.13 | 1.70 |
| Time 2 | 1.00 | 0.15 | 1.27 | 0.13 | -1.39 |
| Time 3 | 1.01 | 0.15 | 1.35 | 0.13 | -1.71 |
| **Mean** | **0.00** | **0.17** | **0.00** | **0.15** | **-0.01** |
| **(SD)** | **(0.80)** | **(0.02)** | **(0.81)** | **(0.02)** | **(1.82)** |

Note: D1.1 represents the barrier calibrations obtained in *Step 1* for September, and D1.2 represents the barrier calibrations obtained in *Step 1* for June. $|z|>2.00$ is considered large enough to indicate unstable perception of barriers across occasions.

Table 3

Rating Scale Step calibrations from *Step 1* and *Step 2*

| Scale Step | F1.1 Logit | F1.2 Logit | F2.1&2 Logit |
|---|---|---|---|
| Unlikely to Minor | -2.05 | -1.45 | -1.69 |
| Minor to Difficult | 0.04 | 0.10 | 0.09 |
| Difficult to Serious | 2.01 | 1.36 | 1.60 |
| **Mean** | **0.00** | **0.00** | **0.00** |
| **(SD)** | **(2.03)** | **(1.41)** | **(1.64)** |

Note: F1.1 represents the rating scale step calibrations obtained in *Step 1* for September, F1.2 represents the scale step calibrations obtained in *Step 1* for June, and F2.1&2 represents the scale step calibrations obtained in *Step 2* for the combined September and June data set (i.e., the common scale).

Table 4

Selected Standardized Differences for Corrected and Uncorrected Teacher Measures

| Teacher | Uncorrected *z* | Corrected *z* | \|Difference\| |
|---|---|---|---|
| 1 | -0.81 | -1.85 | 1.04 |
| 2 | -0.60 | -0.26 | 0.85 |
| 3 | -1.91 | -1.10 | 0.81 |
| 4 | -2.78 | -1.98 | 0.80 |
| 5 | -1.19 | -0.39 | 0.80 |
| 6 | -0.25 | 0.53 | 0.79 |
| 7 | -0.04 | 0.74 | 0.77 |
| 8 | 0.78 | 1.54 | 0.75 |
| 9 | -2.03 | -1.29 | 0.75 |
| 10 | -3.86 | -3.12 | 0.74 |
| 11 | 1.08 | 1.82 | 0.73 |
| 12 | -3.33 | -2.60 | 0.73 |
| 13 | -1.13 | -0.40 | 0.73 |
| 14 | -0.26 | 0.47 | 0.73 |
| 15 | -4.08 | -3.35 | 0.72 |
| 16 | -1.53 | -0.82 | 0.72 |
| 17 | -1.10 | -0.39 | 0.71 |
| 18 | -1.08 | -0.39 | 0.69 |

Table continued.

Table 4 continued.

| | | | |
|---|---|---|---|
| 19 | -3.08 | -2.40 | 0.68 |
| 20 | -1.74 | -2.42 | 0.67 |
| **Mean** | **-0.95** | **-0.60** | **0.46** |
| **(SD)** | **(2.56)** | **(2.47)** | **(0.23)** |

Note: Uncorrected *z* represents the standardized difference in teacher measures obtained for September and June in *Step* 1. Corrected *z* represents the standardized difference in teacher measures obtained in *Step 3* for September and *Step 4* for June. This table shows only the largest 20 absolute differences between uncorrected *z* and corrected *z*. The mean and standard deviation, however, are for the entire group of teachers ($N$=106). Note that a logit difference of 0.95 corresponds to a raw score change of about 0.25 points while a logit change of 0.60 corresponds to a raw score change of about 0.15 points.

Table 5

Standardized Differences for Corrected and Uncorrected Barrier Calibrations

| Barrier | Uncorrected *z* | Corrected *z* | \|Difference\| |
|---|---|---|---|
| Instruction 1 | 0.08 | -0.36 | 0.44 |
| Instruction 2 | -1.69 | -1.96 | 0.28 |
| Instruction 3 | -0.94 | -0.98 | 0.04 |
| Instruction 4 | -2.04 | -1.75 | 0.30 |
| Instruction 5 | -1.20 | -1.52 | 0.32 |
| Instruction 6 | -1.27 | -0.75 | 0.52 |
| People 1 | -3.05 | -2.17 | 0.88 |
| People 2 | -1.94 | -1.19 | 0.75 |
| People 3 | -0.97 | -1.02 | 0.04 |
| People 4 | -1.03 | -1.15 | 0.12 |
| People 5 | 4.62 | 5.31 | 0.69 |
| People 6 | -2.11 | -1.93 | 0.18 |
| People 7 | -0.06 | 0.98 | 1.04 |
| People 8 | 0.84 | 1.84 | 0.99 |
| Resources 1 | -0.62 | -0.12 | 0.50 |
| Resources 2 | 2.57 | 2.28 | 0.29 |
| Resources 3 | 0.93 | 0.65 | 0.28 |
| Resources 4 | 1.73 | 2.16 | 0.43 |

Table continued.

Table 5 continued.

| | | | |
|---|---|---|---|
| Resources 5 | 1.27 | 1.51 | 0.24 |
| Resources 6 | 2.38 | 1.98 | 0.40 |
| Scores 1 | 0.98 | 0.42 | 0.56 |
| Scores 2 | 3.40 | 3.40 | 0.00 |
| Scores 3 | -0.28 | -0.20 | 0.08 |
| Scores 4 | -0.83 | -1.39 | 0.56 |
| Scores 5 | -0.46 | -0.96 | 0.50 |
| Scores 6 | 0.66 | 0.03 | 0.64 |
| Time 1 | 1.70 | 0.32 | 1.38 |
| Time 2 | -1.39 | -2.71 | 1.32 |
| Time 3 | -1.71 | -3.10 | 1.39 |
| **Mean** | **-0.01** | **-0.08** | **0.52** |
| **SD** | **(1.82)** | **(1.92)** | **(0.40)** |

Note: Uncorrected *z* represents the standardized difference in barrier calibrations obtained for September and June in *Step* 1. Corrected *z* represents the standardized difference in barrier calibrations obtained in *Step 3* for September and *Step 5* for June (*N*=29).

## Figure Caption

Figure 1. Steps for creating a frame of reference using Rasch measurement.

Figure 2. Data Layout for *Step 1*

Figure 3. Data Layout for *Step 2*

**STEP 1**

Evaluate the step and item invariance:

$\tau_{time1}$ *versus* $\tau_{time2}$ and $\delta_{time1}$ *versus* $\delta_{time2}$

↓

**STEP 2**

Correct the scale calibrations:

$calibrate\ (\tau_{time1}, \tau_{time2}) = \tau_{corrected}$

↓

**STEP 3**

Benchmark the Time 1 estimates:

$calibrate\ (\delta_{time1} | \tau_{corrected}) = \delta_{time1\ corrected}$

$calibrate\ (\beta_{time1} | \tau_{corrected}) = \beta_{time1\ corrected}$

↓

**STEP 4**

Correct the Time 2 person measures:

$calibrate\ (\beta_{time2} | \beta_{time1\ corrected}, \tau_{corrected}) = \beta_{time2\ corrected}$

↓

**STEP 5**

Correct the Time 2 item difficulties:

$calibrate\ (\delta_{time2} | \delta_{time1\ corrected}, \tau_{corrected}) = \delta_{time2\ corrected}$

$$Time1: \begin{bmatrix} 1 & 1 & X_{1.1.1} \\ 1 & 2 & X_{1.1.2} \\ \vdots & \vdots & \vdots \\ 1 & i & X_{1.1.i} \\ . & . & . \\ . & . & . \\ . & . & . \\ n & 1 & X_{1.n.1} \\ n & 2 & X_{1.n.2} \\ \vdots & \vdots & \vdots \\ n & i & X_{1.n.i} \end{bmatrix} \qquad Time2: \begin{bmatrix} 1 & 1 & X_{2.1.1} \\ 1 & 2 & X_{2.1.2} \\ \vdots & \vdots & \vdots \\ 1 & i & X_{2.1.i} \\ . & . & . \\ . & . & . \\ . & . & . \\ n & 1 & X_{2.n.1} \\ n & 2 & X_{2.n.2} \\ \vdots & \vdots & \vdots \\ n & i & X_{2.n.i} \end{bmatrix}$$

$$\textit{Time}1\ \&\ 2: \begin{bmatrix} 1.1 & 1 & X_{1.1.1} \\ 1.1 & 2 & X_{1.1.2} \\ \vdots & \vdots & \vdots \\ 1.1 & i & X_{1.1.i} \\ 1.2 & 1 & X_{1.2.1} \\ 1.2 & 2 & X_{1.2.2} \\ \vdots & \vdots & \vdots \\ 1.2 & i & X_{1.2.i} \\ . & . & . \\ . & . & . \\ . & . & . \\ n.1 & 1 & X_{n.1.1} \\ n.1 & 2 & X_{n.1.2} \\ \vdots & \vdots & \vdots \\ n.1 & i & X_{n.1.i} \\ n.2 & 1 & X_{n.2.1} \\ n.2 & 2 & X_{n.2.2} \\ \vdots & \vdots & \vdots \\ n.2 & i & X_{n.2.i} \end{bmatrix}$$